ACCEPTED FROM OPEN CALL

ENERGY EFFICIENCY IN MASSIVE MIMO-BASED 5G NETWORKS: OPPORTUNITIES AND CHALLENGES

K. N. R. SURYA VARA PRASAD, EKRAM HOSSAIN, AND VIJAY K. BHARGAVA

ABSTRACT

As we make progress toward the 5G of wireless networks, the bit-per-joule energy efficiency (EE) becomes an important design criterion for sustainable evolution. In this regard, one of the key enablers for 5G is massive multiple-input multiple-output (MIMO) technology, where the BSs are equipped with an excess of antennas to achieve multiple orders of spectral and energy efficiency gains over current LTE networks. Here, we review and present a comprehensive discussion on techniques that further boost the EE gains offered by massive MIMO (MM). We begin with an overview of MM technology and explain how realistic power consumption models should be developed for MM systems. We then review prominent EE-maximization techniques for MM systems and identify a few limitations in the state-of-the-art. Next, we investigate EE-maximization in "hybrid MM systems," where MM operates alongside two other promising 5G technologies: millimeter wave and heterogenous networks. Multiple opportunities open up for achieving larger EE gains than with conventional MM systems because massive MIMO benefits mutually from the co-existence with these 5G technologies. However, such a co-existence also introduces several new design constraints, making EE-maximization non-trivial. A critical analysis of the state-of-the-art EE-maximization techniques for hybrid MM systems allows us to identify several open research problems which, if addressed, will immensely help operators in planning for energy-efficient 5G deployments.

INTRODUCTION

EXPECTATIONS FROM 5G CELLULAR NETWORKS

The information and communication technology (ICT) industry is making rapid progress toward fifth generation (5G) wireless networks, which are expected to integrate almost everything across the globe into the Internet. 5G systems are expected to provide peak data rates up to 20 Gb/s, average data rates greater than 100 Mb/s, and connectivity for a huge number of Internet-of-Things devices per unit area.

Energy consumption becomes a critical concern for 5G networks because the ICT sector already contributes significantly toward the global carbon footprint. In this regard, an important design criterion for 5G networks is bit-per-joule energy efficiency (EE), defined as

$$EE = R/P, \quad (1)$$

where R is the system throughput and P is the power spent in achieving R. The recently proposed massive multiple-input multiple-output (MIMO) technology offers multiple orders of spectral and energy efficiency gains over current LTE technologies, and is therefore a promising enabler for 5G.

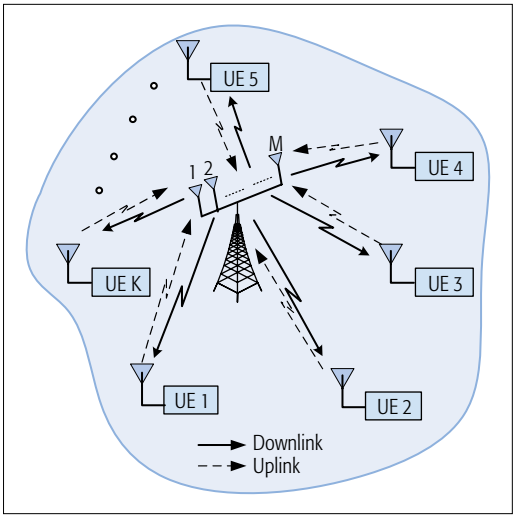

Figure 1. Massive MIMO: a multi-user MIMO technology where K UEs are serviced by a BS with $M >> K$ antennas.

OVERVIEW OF MASSIVE MIMO TECHNOLOGY

Massive MIMO (MM) is a multi-user MIMO (MU-MIMO) technology where K user equipments (UEs) are serviced on the same time-frequency resource by a base station (BS) with M antennas, such that $M \gg K$ (Fig. 1). Deploying a large number of antennas at the BS results in a propagation scenario called *favorable propagation*, where the wireless channel becomes near-deterministic because the BS-to-UE radio links become near-orthogonal to each other [1]. This is because the effects of small-scale fading, intra-cell interference and uncorrelated noise disappear asymptotically in the large M regime. By increasing the size of the system, that is, (M, K), large multiplexing and array gains can be achieved under *favorable propagation*. To understand how, let us consider

K. N. R. Surya Vara Prasad and Vijay K. Bhargava are with University of British Columbia.

Ekram Hossain is with the University of Manitoba.

Digital Object Identifier: 10.1109/MWC.2016.1500374WC

the uplink (UL) and downlink (DL) transmissions in a single-cell MM system. If C_{UL} and C_{DL} are the asymptotic UL and DL Shannon capacities for a flat-fading MU-MIMO channel under *favorable propagation*, we have [2]

$$C_{UL} = \sum_{k=1}^{K} \log_2(1 + p_u M \beta_{k,UL}),$$

$$C_{DL} = \max_{(a_k \geq 0, \sum a_k \leq 1)} \sum_{k=1}^{K} \log_2(1 + p_d M a_k \beta_{k,DL}), \quad (2)$$

where p_u and p_d are the average UL and DL transmit signal to noise ratios (SNRs), $\beta_{k,UL}$ and $\beta_{k,DL}$ represent the large-scale fading coefficients for the k^{th} UE on the UL and DL, respectively, and $\{a_k\}$ is a set of variables that should be optimized to obtain C_{DL}.

When appropriate power control strategies are used to normalize the effect of β_k, (see [3], for example), the UL capacity simplifies to $K\log_2(1 + M_{p_u})$. A similar expression can also be obtained on the DL. This simplification leads us to two important conclusions: we can achieve $O(M)$ array gains, that is, we can reduce the UE transmission power proportionately with M and still achieve the same per-UE throughput as with a single-antenna BS, and $O(K)$ multiplexing gains, that is, we can increase the system throughput proportionately with K by multiplexing parallel streams of data to the UEs. MM systems can also achieve large EE gains over current LTE systems. Before we explain why, a few guidelines are presented on how power expenditure should be modelled for MM systems.

Modelling Power Consumption in Massive MIMO Systems

The sum power consumption P, aggregated over UL and DL transmissions in an MM system, can be modelled as

$$P = P_{PA} + P_C + P_{sys}, \quad (3)$$

where P_{PA} represents the total UL and DL power consumed by the power amplifiers (PAs) at the BS and the UEs, P_C represents the total UL and DL circuit power expenditure, and P_{sys} represents the remaining system-dependant component in P. While P_{PA} accounts for the sum power expenditure on RF transmissions, P_C accounts for the circuit power expenditure on RF chain components, such as filters, mixers, and synthesizers, as well as baseband operations, such as digital up/down conversion, precoding, receiver combining, channel coding/decoding, and channel estimation. P_{sys} accounts for the power expenditure on site-specific and architecture-specific factors, such as BS architecture, power supply, cooling system, backhaul, and other control equipment. P_{sys} will play an important role in characterizing EE for 5G networks because several BS and UE types will simultaneously operate under an architecture with multiple cell sizes and access technologies. Different MM systems generally exhibit different P_{PA}, P_C, and P_{sys} values, depending on the size of the system and the type of hardware components used, including transceivers, PAs, RF chains, backhaul equipment, and switches. Example values can be found in [3, 4] and references therein.

Note that P_C in MM networks should not be modelled, as per conventional practice, as a constant term that is independent of (M, K), because the hardware requirements and the number of circuit operations in the system grow with M and K. For example, with the conventional *one RF chain per antenna* design used in current LTE systems, the number of RF chains at the BS and the UEs grows affinely with M and K, respectively. Additionally, computational requirements for various baseband operations are functions in (M, K). For example, as illustrated in [3], $O(MK^2)$ computations are required for zero-forcing (ZF) precoding as well as minimum mean squared error (MMSE) channel estimation. Therefore, P_C should be treated as a function in (M, K).

Energy Efficiency Aspects of Massive MIMO

When compared to conventional MU-MIMO systems under LTE, MM systems can achieve large EE gains in two major ways, both based on increasing the size of the system, that is, (M, K). First, for a given system throughput, transmission power of the UEs in MM systems can be reduced significantly by increasing M (Eq. 2) well beyond the maximum limit of eight *antennas per BS* in current LTE systems. Second, by increasing K, large throughput gains can be achieved in MM systems (Eq. 2). These gains can be achieved at low levels of circuit power expenditure because simple linear processing techniques, such as maximal-ratio combining (MRC) on the UL and maximal-ratio transmission (MRT) on the DL, can achieve near-optimal throughput performance [1]. The resulting EE levels are generally much higher than in conventional MU-MIMO systems because the latter systems employ complex signal processing techniques, such as maximum likelihood (ML) detection on the UL [5] and dirty paper coding (DPC) on the DL, to achieve optimal capacities. Such complex techniques can consume prohibitively large computational power when the size of the system is increased. For example, computational power for ML detection increases exponentially with K [5].

However, increasing (M, K) in MM systems does not always ensure improvements in EE, because the circuit power expenditure, that is, P_C, also increases with (M, K). Consequently, when the UE transmission powers are reduced by increasing M, we can achieve EE improvements only if the reduction in UE transmission power dominates the resulting increase in P_C. Similarly, when the system throughput is increased by increasing K, EE improvements can be achieved only if the increase in throughput dominates the resulting increase in power expenditure. See [3, 4] for examples on how M and K in MM systems can be optimized to achieve large EE gains over current LTE systems.

Designing Energy-Efficient Massive MIMO Systems

As we can observe from Eq. 1, EE of an MM system can be maximized by achieving near-optimal throughput performance at low power consumption levels. Based on this analogy, a number of research directions have been pursued for the

The recently proposed massive multiple-input multiple-output (MIMO) technology offers multiple orders of spectral and energy efficiency gains over current LTE technologies, and is therefore, a promising enabler for 5G.

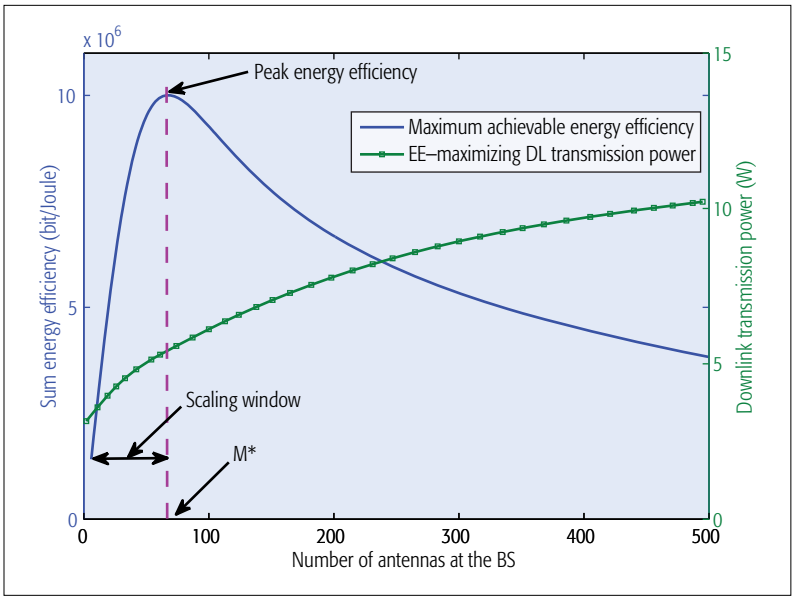

Figure 2. Maximum achievable sum EE vs. number of BS antennas in an MM system. As illustrated, there exists a scaling window, over which the sum EE can be improved by increasing the number of BS antennas.

design of energy-efficient MM networks. A few methods devise low-complexity algorithms for BS operations such as multi-user detection, precoding, and user scheduling, so as to minimize power expenditure in the system. A few other methods, such as transceiver redesign, antenna selection, and power amplifier dimensioning, focus on improving resource utilization so as to relax hardware requirements, and hence power expenditure in the system. This section reviews some of the most prominent EE-maximization techniques for MM systems and identifies a few open research problems.

Low-Complexity BS Operations

Due to favorable propagation in the large M regime, simple linear processing techniques, such as MRC on the UL and MRT on the DL, and simple user scheduling algorithms, such as random and round-robin scheduling, achieve near-optimal throughputs [1]. These simplifications keep the circuit power expenditure low, thus yielding significant EE gains over conventional MU-MIMO systems with computationally intensive signal processing schemes, such as ML detection and DPC, and complex user scheduling algorithms, such as random beamforming and semi-orthogonal user selection.

One of the major research challenges for massive MIMO is the design of low-overhead frequency division duplex (FDD) precoders. Unlike in time division duplex (TDD) systems, precoders in FDD systems cannot exploit channel reciprocity to estimate DL channels based on UL channels, because the UL and DL communications occur on separate frequency bands. FDD precoders cannot also rely on pilot signalling and feedback from the UEs because this consumes at least (M + K) symbols per coherence interval, making such signalling and feedback mechanisms impractical for high mobility scenarios. A few low-overhead FDD precoders, which assume channel sparsity for overhead reduction, have been proposed recently [6], but such precoders are limited to high frequency bands, such as millimeter wave, where channel sparsity assumptions are valid. Worldwide, since there are many more licenses for FDD than TDD, progress on low-overhead FDD precoders will promote wider acceptance of MM as a future technology.

Scale the Number of BS Antennas

When the number of antennas at the BS is increased, the system throughput R can be improved because higher multiplexing gains are achievable. However, observe from our discussions above that the circuit power P_C also increases with M. Nevertheless, increasing M can still be an energy-efficient strategy for MM networks if we increase R sufficiently that it dominates the increase in P_C. Although not obvious from initial observations, this can be done by increasing the DL transmission power over a certain scaling window. To understand why, let us study how DL transmission power can be optimized for maximizing energy efficiency.

As we can observe from Eqs. 1–3, the system EE is a non-linear function in the DL transmission power because the EE metric takes a fractional form where the numerator, that is, the system throughput R, and the denominator, that is, the power consumption P, are both functions in the DL transmission power. Therefore, to optimize the DL transmission power for EE-maximization, we can use standard non-linear optimization methods, such as gradient descent [7]. Optimal DL transmission powers would be different for different (M, K) because the system EE depends on (M, K) as well. Using this methodology, in Fig. 2 we plot the maximum achieved EE and the corresponding DL transmission powers when M is increased from 5 to 500. For the simulation, we assume K = 30, MRT precoding, and use the power consumption model given in [3].

From Fig. 2 we clearly observe that there exists a scaling window within which the system EE can be increased by simultaneously increasing M and the DL transmission power. The scaling window is governed by a certain threshold M^* on the number of BS antennas, beyond which R approaches near-optimal bounds but P_C continues to grow unboundedly with M. As a result, we observe that the system EE attains a peak level at M^* and decreases gradually with M, even if we increase the DL transmission power. Note that the scaling window can be expanded by reducing the RF chain requirements at the BS because this results in reduced P_C levels. Alternatively, if the transceiver design ensures very low circuit power expenditure, it may also be possible to achieve improvements in the system EE by reducing the DL transmission power with M. However, such a scenario relies heavily on the future of energy-aware transceiver design.

Minimize PA Power Losses

Significant EE gains can also be achieved by minimizing PA power losses. To understand how, let us consider DL transmissions in an MM system. The total power expenditure on PAs at the BS can be obtained as $P_{PA} = P_{in}/\eta$, where P_{in} is the input power to PAs and η is a measure of PA power losses (also referred to as the PA efficiency). For

traditional PAs in current LTE systems, η depends on the PA output power p and is given by

$$\eta = \eta_{\max}\sqrt{p / p_{\max}}$$

in [8], where $\eta_{\max}$ is the maximum PA efficiency and $p_{\max}$ is the maximum PA output power. The expression for η provides two important conclusions:

- η is maximum, that is, $\eta = \eta_{\max}$ when $p = p_{\max}$
- $\eta < \eta_{\max}$ when $p < p_{\max}$.

Average p in current LTE systems can be much lower than $p_{\max}$ because OFDM waveforms have a high peak-to-average-power ratio (PAPR) requirement [9]. Consequently, average PA efficiencies, that is, η, can be much smaller than $\eta_{\max}$. Smaller η results in higher P_{PA}, and therefore in smaller EE (Eqns. 1, 3). EE gains from minimizing PA power losses, that is, from maximizing η, can be significant because η can be as low as 5 percent in LTE [10].

To minimize PA power losses, several PAPR reduction techniques, such as proposed in [9], can be attempted. In addition, few low-PAPR non-orthogonal waveforms, such as the recently proposed single carrier modulation (SCM) [11], can be used. However, designing appropriate non-orthogonal waveforms continues to be a major research challenge because most of the recently proposed waveforms suffer from practical limitations, such as long filter lengths and complex receiver techniques [11]. Alternatively, PA linearity requirements can also be relaxed using constant envelope signals, but generating such signals is an unresolved challenge as of this writing.

Minimize RF Chain Requirements at the BS

Conventionally, MIMO precoding and beamforming are performed digitally in the baseband. Since digital processing requires dedicated baseband and RF chain components for each antenna element, BS transceivers conventionally adopt a one *RF chain per antenna* design. Such a design results in significant circuit power consumption in the MM regime because the number of RF chains at the BS increases affinely with M. Therefore, minimizing RF chain requirements at the BS is an attractive strategy to improve EE in MM networks. Prominent techniques to reduce RF chain requirements include hybrid precoding, antenna selection, and transceiver redesign. Hybrid precoding techniques are generally built on channel sparsity assumptions, and hence are discussed below in the context of millimeter wave systems.

Antenna Selection: Antenna selection is a signal processing technique that improves throughput in a system while simultaneously reducing the number of RF chains at the BS [12]. Basically, a subset comprising N out of the M BS antennas is selected based on a predefined selection criterion, for example, to maximize throughput, SNR, or EE. Antennas in the selected subset are connected to RF chains for further processing. Since the number of RF chains is reduced from M to N, circuit power expenditure in the system is reduced.

In Fig. 3, we propose a guideline to design traffic-adaptive antenna selection methods for energy efficiency in a massive MIMO system. To study how the system EE varies with traffic demands, we plot the maximum achievable EE as shown in Fig. 2, but for K values ranging from 6 to 50.

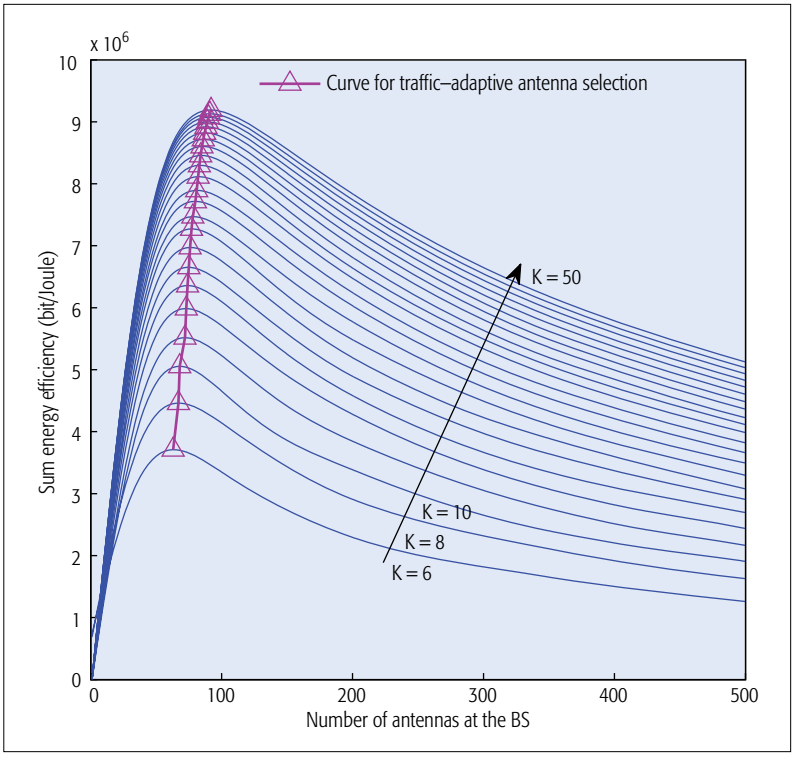

Figure 3. Guideline to design traffic-adaptive antenna selection schemes for EE-maximization in massive MIMO systems.

To obtain the maximum EE values for each K, we make the same assumptions and follow the same optimization procedure as done in Fig. 2. When K is increased, the system EE increases because higher throughputs can be achieved if more UEs are scheduled. However, since an increase in K also results in increased P_C, we observe that the per-unit increase in EE decreases with K. Additionally, we observe that the maximum achievable EE curves attain different peak levels when M is optimized for different K values. Specifically, when K is increased from 6 to 50, optimal M increases from 63 to 92 and the peak EE increases from 3.6 to 9.2 Mb/Joule.

Assuming that the maximum expected traffic demand corresponds to K = 50, it is natural to expect that the BS is deployed with a total of 92 antennas. For such a BS, the above variation in optimal M, that is, from 63 to 92, amounts to about 30 percent of the total number of deployed antennas. Therefore, although it may seem otherwise from an initial observation, the demonstrated variation in optimal M with K is indeed significant. Based on this observation, we can design traffic-adaptive antenna selection methods, where the BS can dynamically activate a subset of its antennas with changing traffic demands in the system. The BS can navigate along the antenna selection curve in Fig. 3 for sustained operation at peak EE levels, even if traffic demands vary with time. It is important to note that the curve for traffic-adaptive antenna selection shown in Fig. 3 does not necessarily need to be a monotonically increasing function in K because the power consumption parameters assumed in our simulation (and given in [3]) may not always be accurate. Real-time data from network operators should be utilized in developing accurate power consump-

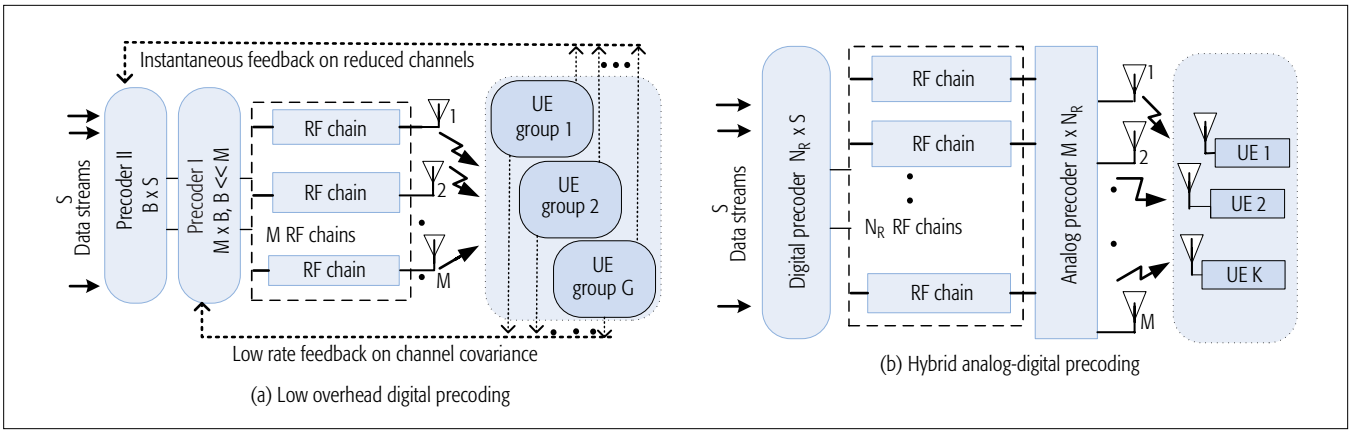

Figure 4. Novel precoding techniques for mmWave massive MIMO systems: a) Two-stage digital precoding techniques can reduce CSI overhead by forming UE groups, each group comprising UEs with the same covariance eigenspace; b) Hybrid analog-digital precoding techniques can exploit channel sparsity to reduce RF chain requirements at the BS.

tion models, and thereby in devising appropriate antenna selection curves.

Current literature on antenna selection for massive MIMO is mostly confined to simple single cell scenarios (see [12] and references therein). Performance trade-offs introduced by design limitations, such as CSI availability, pilot contamination, and antenna correlation, are not clearly understood.

Redesign Transceiver Architecture: An alternative strategy to reduce RF chain requirements at the BS is to redesign the BS transceiver architecture. In this direction, a few *single RF chain transceivers* have been recently designed, although at the cost of some serious practical limitations. For example, the electronically steerable parasitic antenna array proposed in [13] operates with a single RF chain, but supports a limited set of modulation schemes and requires almost twice the number of antennas than in conventional transceivers. Similarly, [14] proposes a *single RF chain transmitter* based on a two-port matching network, but the transceiver performance is subject to power losses in the matching network and mutual coupling in the antenna array. Consequently, although transceiver redesign offers great promise to improve EE in MM networks, current literature cannot be considered complete. Further research is required on addressing several design issues and on overcoming any implementation challenges thereof.

Energy Efficiency in Hybrid Massive MIMO Systems

So far, we have explored multiple opportunities for EE-maximization in conventional MM systems, that is, wireless systems where MM is the only enabling 5G technology. Several other technologies, such as millimeter wave (mmWave), heterogenous networks (HetNets), energy harvesting (EH), full duplex, and cloud-based radio access, are also emerging as potential enablers for 5G. Each of these offers its own unique set of performance benefits: mmWave operations offer throughput enhancements via larger transmission bandwidths; EH allows for low battery power expenditure via renewable energy usage; and HetNets offer large throughput gains via network densification. Therefore, it is natural to expect that future 5G architectures will host wireless systems enabled by MM as well as other emerging 5G technologies. We refer to such systems as "hybrid MM systems" and investigate potential opportunities for EE-maximization in this section.

We explain how MM benefits mutually from two promising 5G technologies, namely mmWave and HetNets, and proceed to study EE-maximization in the corresponding hybrid MM systems. Some of the unique properties of these systems are studied to understand why new opportunities and design constraints emerge from an EE perspective. A critical review of the state-of-the-art allows us to identify several open research problems which, if pursued, will immensely help operators in touching unexplored avenues and in understanding crucial design considerations for hybrid MM deployments.

Millimeter Wave (mmWave)-Based MM Systems

The mmWave spectrum ranging from 30 GHz to 300 GHz is now being investigated for 5G operations because the sub-3GHz bands have become overcrowded and there is a need for additional spectrum to accommodate future traffic demands. By moving to the mmWave spectrum, significant throughput gains and latency reductions can be achieved because large bandwidths of the order of multiple GHz are available; bandwidths up to 7 GHz are available in the 60 GHz band. Typically, mmWave channels exhibit huge reflection and absorption losses, poor diffraction, and low channel coherence times. As a result, when compared to sub-3GHz bands, mmWave channels experience much higher channel correlation, signal attenuation, and sensitivity to blockage [6, 15].

Benefits from Co-Existence: Massive MIMO implicitly offers the highly directional and adaptive transmissions required to improve signal strength and suppress interference in the blockage-sensitive environments at mmWave bands. On the other hand, mmWave makes massive MIMO realizable because the small wavelengths at mmWave frequencies allow a large number of antennas to be fit into very small form factors [6], and the near-LOS channels in mmWave MM networks

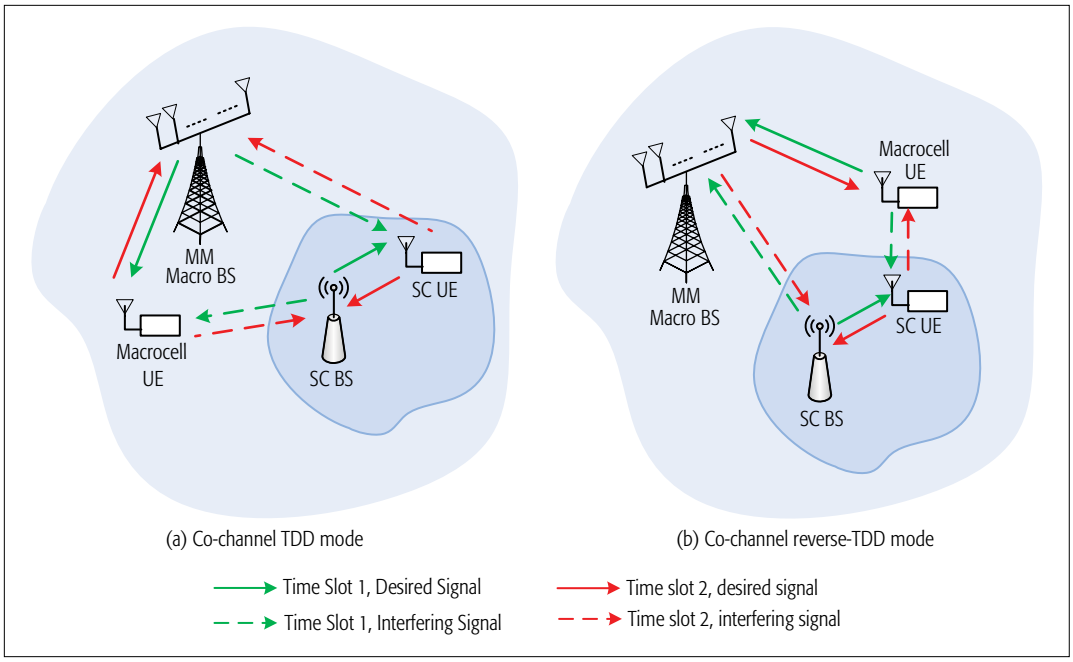

Figure 5. Co-channel TDD and co-channel reverse TDD deployment modes for massive MIMO HetNets. Throughput gains are achieved because the available spectrum is fully utilized in the macro-tier as well as the SC tier.

Typically, mmWave channels exhibit huge reflection and absorption losses, poor diffraction, and low channel coherence times. As a result, when compared to sub-3GHz bands, mmWave channels experience much higher channel correlation, signal attenuation, and sensitivity to blockage.

can be estimated using direction of arrival (DoA) of the incident waves at the BS, thus potentially eliminating the need for pilot reuse and the resulting pilot contamination [15].

There are two major differences between mmWave-based MM systems and traditional mmWave systems. First, most traditional mmWave systems achieve directional transmissions using analog phased arrays with a limited number of antennas [16]. In contrast, BSs in mmWave MM systems employ digital beamforming and spatial multiplexing with a much larger number of antennas. Second, due to strong signal attenuation and blockage, traditional mmWave systems are limited to point-to-point or short-range indoor services, such as radar, backhaul, local area networks (see IEEE 802.11ad), and personal area networks (see 802.15.3c) [16]). In contrast, thanks to beamforming with a large number of antennas, mmWave MM systems can also be used for longer-range cellular services with simultaneous multi-user transmissions.

Opportunities for Energy-Efficient Design: Low-complexity channel equalization techniques are sufficient for mmWave MM networks because the narrow directional beams and the near-LOS propagation eliminate much of the multipath. For the same reason, mmWave MM BSs can support point-to-multipoint wireless backhaul operations, thus becoming a cost-effective alternative to fiber backhaul. In addition, the FDD mode of operation, which incurs large CSI overhead and is therefore impractical at the sub-3GHz bands, becomes realizable at mmWave bands because sparsity in mmWave channels can be exploited to derive low-overhead multi-stage precoding techniques. To explain how, we consider the two-stage precoding technique proposed in [6] and illustrate the underlying idea in Fig. 4a.

The two-stage digital precoding technique in Fig. 4a exploits channel sparsity to partition UEs into different groups, each group comprising UEs with approximately the same channel covariance eigenspace, such that the covariance eigenspaces of different UE groups are near-orthogonal to each other. To understand how the CSI overhead is reduced, let us first denote r as the rank of channel covariance matrix and S, where ($S \leq K$), as the number of independent streams to be transmitted to the UEs. Precoder I exploits the near-orthogonality of covariance eigenspaces to reduce the channel dimensionality from ($M \times K$) to ($B \times S$), where B ($S \leq B < r$) is an optimization parameter to regulate intergroup interference in the system. A low-rate feedback mechanism is sufficient to update precoder I because it depends only on the channel covariance, which typically varies very slowly when compared to the channel coherence time. Precoder II employs simple linear precoding techniques on the effective ($B \times S$) channel so as to extract multiplexing gains within each UE group. To update precoder II, the BS should acquire instantaneous CSI of the effective ($B \times S$) channel during each coherence interval. Observe that the CSI overhead will still be significantly lower than in conventional FDD systems because the overhead comes predominantly from estimating reduced-dimensional channels.

Sparsity in mmWave channels can also be exploited to design hybrid analog-digital beamforming techniques that relax RF chain requirements in the system [17]. As an example, we consider the hybrid precoding technique proposed in [17] and illustrate the underlying idea in Fig. 4b The hybrid precoding technique in Fig. 4b reduces the number of RF chains from M to N_R, where $S \leq K$, $S \leq N_R \leq M$. The analog precoder applies phase-only control to extract large array gains and to reduce the channel dimensionality from $M \times K$ to $N_R \times S$. The digital precoder applies simple linear precoding techniques on the effective $N_R \times S$ channel to extract multiplexing gains. RF chain requirements are reduced because

Due to smaller coverage areas, SCs fail to ensure seamless connectivity and quality of service (QoS) to UEs which are highly mobile. This limitation can be overcome by designing a two-tier MM HetNet, wherein a macro cell tier formed by the MM BSs is overlaid with an SC tier formed by small cells, such as pico cells and femto cells.

the digital precoders operate only on the effective low-dimensional channel.

Challenges and Open Problems: Despite clear evidence that multi-stage digital precoding techniques, as shown in Fig. 4a, can be designed to reduce training overhead in mmWave MM systems, such techniques have only been studied to a limited degree (see [6] for an example). Trade-offs introduced by pilot contamination are not clearly understood. Missing in the existing literature are studies that optimize the interference mitigation parameter B for energy efficiency. Other open problems include optimizing user grouping, covariance tracking, and inter cell interference mitigation for energy efficiency. Similar is the situation with hybrid analog-digital beamforming techniques that relax RF chain requirements at the BS. These techniques are invaluable for mmWave operations because mixed signal components in the RF chain, particularly the high resolution analog to digital converters (ADCs), consume unacceptably large amounts of power under large-bandwidth operations. Notice that the analog precoding phase introduces several new constraints on the transceiver design, such as limited precision for phase control, limited number of phase shifts, and limited ADC resolution. Existing literature does not discuss the EE trade-offs introduced by these constraints, leaving a huge scope for further research.

Another major bottleneck in the realization of energy-efficient mmWave MM systems is the hardware design. Silicon-based CMOS technologies provide a simple and cost-effective means to integrate several mmWave antennas with necessary analog and digital circuitry onto a single package. However, the high frequency and large-bandwidth operations in the mmWave regime impose several constraints on the design of transceiver components. For example, high substrate absorption losses and high noise power levels become roadblocks to the design and integration of highly directional antennas into CMOS packages. In addition, improper isolation between active on-chip components can result in self-jamming and signal distortion. Transceivers that address all of these design complications have not been fabricated as of this writing.

MM-Based Heterogenous Networks

Dense heterogenous networks (HetNets), where spectrum utilization is maximized by decreasing the cell size and increasing the number of small cells (SCs) per unit area, offer a promising approach to satisfy the traffic demands expected in 5G. In terms of EE, HetNets are a superior alternative to massive MIMO because the power consumption per small cell access point (SCA) is generally low, SCAs can be opportunistically turned on/off depending on traffic demand, and high throughput gains can be achieved by intelligently offloading traffic between outdoor and indoor SCs. Moreover, when M SCs are deployed per unit area, the average BS-to-UE distance is reduced by $M^{(1/2)}$ Therefore, if γ is the path loss exponent, we can reduce the UE transmission powers proportionately with $M^{(\gamma/2)}$ and still achieve the same per-UE throughput as before densification. In other words, we can achieve $O(M^{(\gamma/2)})$ array gains. Since $\gamma > 2$ for most propagation conditions, these array gains are generally larger than the $O(M)$ gains offered by conventional MM systems.

Benefits from Co-Existence: Due to smaller coverage areas, SCs fail to ensure seamless connectivity and quality of service (QoS) to UEs that are highly mobile. This limitation can be overcome by designing a two-tier MM HetNet, wherein a macro cell tier formed by the MM BSs is overlaid with an SC tier formed by small cells, such as pico cells and femto cells. The macro cell tier ensures uniform service coverage and supports highly mobile UEs, while the SC tier caters to local indoor and outdoor capacity requirements. Clearly, such an architecture can simultaneously extract the $O(M^{(\gamma/2)})$ array gains offered by HetNets and the $O(K)$ multiplexing gains offered by MM. In addition, since the macro tier hosts a large number of antennas, few antennas can be reserved for wireless backhaul to the SC tier. Interference coordination in MM HetNets can be analyzed by using simple tools from random matrix theory [18–20]. This is highly beneficial because tools from stochastic geometry, which are used to study interference coordination in single antenna HetNets, cannot be easily applied to MM HetNets due to the introduction of cross-tier statistical dependencies.

Opportunities for Energy-Efficient Design: EE in MM HetNets can be improved by combining the EE-maximization techniques discussed above with few EE-maximization techniques for HetNets, such as BS sleeping, cell zooming, cell association, and coordinated multi-point transmission (CoMP). Several other EE-maximization techniques can be designed by jointly exploiting the properties of MM and HetNet technologies. For example, MM HetNets can use low-complexity multi-flow beamforming techniques to jointly coordinate interference among UEs in both macro and SC tiers. Such techniques are known to drastically reduce hardware requirements at the MM BSs. For example, [18] shows that the number of MM BS antennas can be reduced by more than 50 percent if a few single antenna SCs are overlaid on the MM cell.

In addition, co-channel TDD deployment modes, where the available spectrum is fully utilized in both macro and SC tiers, can be attempted. For illustration, let us consider two example scenarios proposed in [20], namely the co-channel TDD (co-TDD) and the co-channel reverse TDD (co-RTDD) modes. The underlying ideas are illustrated in Fig. 5. In the co-TDD mode, the macro and SC tiers are time-synchronized to simultaneously transmit in the UL or the DL. In the co-RTDD mode, the order of UL and DL transmissions are reversed in one of the tiers, that is, the macro tier operates in the DL when the SC tier operates in the UL, and vice versa. Since the entire spectrum is utilized in both the tiers, simultaneous and uncoordinated transmissions can introduce significant inter-tier and intra-tier interference. Fortunately, the BSs in the macro and SC tiers can not only estimate the channels to their intended UEs, but also the covariance of interfering signals. As a result, channel reciprocity can be exploited to design precoding vectors that sacrifice certain degrees of freedom (DoFs) on the DL so as to blank out the strongest interference subspace [19, 20]. When such *spatial blanking*

techniques are used and the number of sacrificed DoFs are optimized (see [19] for an example), significant throughput gains can be achieved in the SC tier at the cost of a negligible throughput loss in the macro tier.

The co-TDD and co-RTDD modes exhibit conflicting properties, leading to some interesting trade-offs during the design of energy-efficient MM HetNets. For example, the quality of interference estimation and the ability to reject interference can be considerably different because the interfering signals are radically different. Co-RTDD renders higher interference estimation accuracy than co-TDD because the interferer channels are quasi-static in co-RTDD, due to fixed locations of the MM and SC BSs, but are dynamically varying in co-TDD, due to moving UEs. Consequently, when in the macro-tier UL, co-RTDD can attempt *spatial blanking* to achieve higher throughput gains than co-TDD [19]. On the other hand, when in the macro-tier DL, co-RTDD offers lower throughput gains than co-TDD because co-RTDD renders lower interference rejection. This is in turn because the SCs have many fewer antennas than the MM BSs and hence, sacrificing DoFs at the SCs may not reduce the cross-tier interference significantly.

Challenges and Open Problems: Several challenges continue to roadblock the design of energy-efficient MM HetNets. For example, most studies on *spatial blanking* (see [19, 20] and references therein) attempt channel covariance estimation and precoding based on a wide-sense stationarity assumption on the channel process. Such an assumption is valid only locally and is susceptible to UE mobility. Therefore, novel channel tracking algorithms should be developed to adaptively learn and update the estimated interference subspace according to the non-stationary time-varying effects in the system. Also, most studies on *spatial blanking* (see [20] and references therein) focus on simplistic UE distribution scenarios with either isolated UEs or hotspots. In practice, HetNets would experience asymmetric traffic loads coming from a combination of hotspots and isolated UEs. Therefore, low-complexity interference coordination strategies should be designed to allow efficient spatial resource sharing between hotspots and isolated UEs. Additionally, as discussed earlier, there is no clear winner among co-TDD and co-RTDD. This calls for the design of innovative co-channel deployment modes, which can simultaneously reap the benefits and overcome the limitations of co-TDD and co-RTDD. Appropriate pilot assignment methods should be developed to contain pilot contamination, which can be particularly severe in co-channel deployments. Load balancing in MM HetNets is another largely unexplored subject. Resource efficient inter-tier offloading techniques based on load-adaptive cell zooming, dynamic antenna activation, and mobility-aware handover, should be designed under practical constraints such as limited backhaul and load asymmetry.

Conclusion

Massive multiple-input multiple-output (MIMO) is a promising technology for sustainable evolution toward 5G because it offers multiple orders of spectral and energy efficiency (EE) gains over current LTE technologies. This article has explored several opportunities to boost the EE gains offered by massive MIMO (MM) systems. Standard EE-maximization techniques for conventional MM systems, such as scaling the number of BS antennas, implementing low-complexity operations at the BS, minimizing power amplifier losses, and minimizing RF chain requirements, were briefly reviewed and a few open research problems were identified. This article has also investigated several new opportunities for EE-maximization in "hybrid MM systems," where MM operates alongside other emerging 5G technologies, such as millimeter wave and heterogenous networks. Mutual benefits arising from the co-existence of these 5G technologies were analyzed to understand why hybrid MM systems have an enormous potential to achieve larger EE gains than conventional MM systems. A critical review of the state-of-the-art on the design of energy-efficient hybrid MM systems allowed us to identify several open research problems for future work. Despite being largely unexplored, hybrid MM systems are promising for future 5G deployments because there is strong evidence that these systems have the potential to meet the energy efficiency demands expected in 5G cellular networks.

Despite being largely unexplored, hybrid MM systems are promising for future 5G deployments because there is strong evidence that these systems have the potential to meet the energy efficiency demands expected in 5G cellular networks.

Acknowledgment

The work was supported by a CRD grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] T. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, Nov. 2010, pp. 3590–3600.
[2] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Sig. Process. Mag.*, vol. 30, no. 1, Jan. 2013, pp. 40–60.
[3] E. Bjornson *et al.*, "Optimal Design of Energy Efficient Multi-User MIMO Systems: Is Massive MIMO the Answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, June 2015, pp. 3059–75.
[4] K. N. R. Surya Vara Prasad and V. K. Bhargava, "Resource Optimization for Energy Efficiency in Multi-Cell Massive MIMO with MRC Detectors," *Proc. IEEE WCNC*, Doha, Qatar, 2016, pp. 1–6.
[5] S. Verdú, " Computational Complexity of Optimum Multiuser Detection," *Algorithmica*, vol. 4, 1989, pp. 303–12.
[6] A. Adhikary *et al.*, "Joint Spatial Division and Multiplexing for mmwave Channels," *IEEE JSAC*, vol. 32, no. 6, Jun. 2014, pp. 1239–55.
[7] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, New York, Springer-Verlag, 1999.
[8] M. M. A. Hossain and R. Jantti, "Impact of Efficient Power Amplifiers in Wireless Access," *Proc. IEEE GreenCom*, 2011, pp. 36–40.
[9] T. Jiang and Y. Wu, "An Overview: Peak-to-Average Power Ratio Reduction Techniques for OFDM Signals," *IEEE Trans. Broadcast.*, vol. 54, no. 2, June 2008, pp. 257–68.
[10] EARTH, 2010, Energy Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown; available: https://www.ict-earth.eu/publications/deliverables/deliverables.html.
[11] G. Wunder *et al.*,"5GNOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 97–105.
[12] Z. Zhou *et al.*, "Energy-Efficient Antenna Selection and Power Allocation for Large-Scale Multiple Antenna Systems with Hybrid Energy Supply," *Proc. IEEE GLOBECOM*, Austin, TX, 2014, pp. 2574–79.
[13] A. Kalis *et al.*, "A Novel Approach to MIMO Transmission Using a Single RF Front End," *IEEE JSAC*, vol. 26, no. 6, Aug. 2008, pp. 972–80.
[14] M. A. Sedaghat *et al.*, "A Novel Single-RF Transmitter for Massive MIMO," *Proc. VDE Int'l. ITG WSA*, Erlangen, 2014, pp. 1–8.

[15] L. Liu *et al.*,"DoA Estimation and Achievable Rate Analysis for 3D Millimeter Wave Massive MIMO Systems," *Proc. IEEE Int'l. Workshop SPAWC*, Toronto, 2014, pp. 6–10.
[16] Y. Niu *et al.*, "A Survey of Millimeter Wave Communications (mmWave) for 5G: Opportunities and Challenges," *J. Wireless Netw.*, vol. 21, no. 8, pp. 2657–76. Springerlink; DOI 10.1007/s11276-015-0942-z.
[17] A. Alkhateeb *et al.*, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," IEEE *J. Sel. Topics Signal Process.*, vol. 8, no. 5, Oct. 2014, pp. 831–46.
[18] E. Bjornson *et al.*, "Massive MIMO and Small Cells: Improving Energy Efficiency by Optimal Soft-Cell Coordination," *Proc. IEEE Int. Conf. Telecommun.*, Casablanca, 2013, pp. 1–5.
[19] J. Hoydis *et al.*, "Making Smart Use of Excess Antennas: Massive MIMO, Small Cells, and TDD," *Bell Labs Tech. J.*, vol. 18, Sept. 2013, pp. 5–21.
[20] A. Adhikary *et al.*, "Massive MIMO Meets HetNet: Interference Coordination through Spatial Blanking," *IEEE JSAC*, vol. 33, no. 6, June 2015, pp. 1171–86.

Biographies

K. N. R. Surya Vara Prasad is a Ph.D. student in electrical and computer engineering at the University of British Columbia (UBC), Vancouver, Canada. He obtained his M.A.Sc. degree in electrical and computer engineering from UBC Vancouver, in 2016, and his B.Tech degree in electrical engineering from the Indian Institute of Technology (IIT) Bhubaneswar, India, in 2012. His current research focus is on resource allocation, distributed optimization, and applications of machine learning in wireless communication networks. He was a recipient of the Best Demo & Exhibits Award at IEEE COMSNETS 2014.

Ekram Hossain [F'15] is a professor in the Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Canada. He is a member (Class of 2016) of the College of the Royal Society of Canada. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He has authored/edited several books in these areas (http://home.cc.umanitoba.ca/~hossaina). He was elevated to an IEEE Fellow "for contributions to spectrum management and resource allocation in cognitive and cellular radio networks." Currently he serves as an editor for *IEEE Wireless Communications*. Also, he is a member of the IEEE Press Editorial Board. Previously, he served as the editor-in-chief of *IEEE Communications Surveys and Tutorials* from 2012–2016, and as the area editor for the *IEEE Transactions on Wireless Communications* in the area of "Resource Management and Multiple Access" from 2009-2011, as an editor for the *IEEE Transactions on Mobile Computing* from 2007-2012, and as an Editor for the *IEEE Journal on Selected Areas in Communications — Cognitive Radio Series* from 2011-2014. He has won several research awards, including the IEEE Vehicular Technology Conference (VTC 2016–Fall) Best Student Paper Award as a co-author, the IEEE Communications Society Transmission, Access, and Optical Systems (TAOS) Technical Committee's Best Paper Award at IEEE Globecom 2015, a University of Manitoba Merit Award in 2010, 2013, 2014, and 2015 (for research and scholarly activities), the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 (WCNC'12) Best Paper Award. He was a Distinguished Lecturer of the IEEE Communications Society (2012-2015). He is a registered professional engineer in the province of Manitoba, Canada.

Vijay Bhargava [S'70, M'74, SM'82, F'92, LF'13] obtained BASc, MASc, and Ph.D. degrees from Queen's University at Kingston in 1970, 1972, and 1974 respectively. He is a professor in the Department of Electrical and Computer Engineering at the University of British Columbia in Vancouver, where he served as department head from 2003 to 2008. Previously he was with the University of Victoria (1984-2003), Concordia University (1976-1984), the University of Waterloo (1976), and the Indian Institute of Science (1974-1975). He has held visiting appointments at Ecole Polytechnique de Montreal, NTT Research Lab, Tokyo Institute of Technology, the University of Indonesia, the Hong Kong University of Science and Technology, and Tohoku University. He is an honorary professor at UESTC, Chengdu, and a Gandhi distinguished professor at IIT Bombay. He is in the Institute for Scientific Information (ISI) Highly Cited list. He served as the founder and president of Binary Communications Inc. (1983–2000). He is a co-author (with D. Haccoun, R. Matyas, and P. Nuspl) of *Digital Communications by Satellite* (New York: Wiley: 1981), which was translated into Chinese and Japanese. He is a co-editor (with S. Wicker) of *Reed Solomon Codes and their Applications* (IEEE Press: 1994), a co-editor (with H.V. Poor, V. Tarokh, and S. Yoon) of *Communications, Information and Network Security* (Kluwer: 2003), a co-editor (with E. Hossain) of *Cognitive Wireless Communication Networks* (Springer: 2007), a co-editor (with E. Hossain and D. I. Kim) of *Cooperative Wireless Communications Networks* (Cambridge University Press: 2011), a co-editor (with E. Hossain and G. Fettweis) of *Green Radio Communications Networks* (Cambridge University Press: 2012), and a co-editor (with D. Niyato, E. Hossain, D. I. Kim, and L. Shafai) of *Wireless-Powered Communication Networks* (Cambridge University Press: 2016). He is a Fellow of the IEEE, the Royal Society of Canada, Canadian Academy of Engineering, and the Engineering Institute of Canada. He is a Foreign Fellow of the National Academy of Engineering (India) and has served as a distinguished visiting fellow of the Royal Academy of Engineering (U.K.). He is a recipient of the 2015 Killam prize in engineering awarded by the Canada Council, and a Humboldt Research Award from the Alexander von Humboldt Stiftung. He has served as the editor-in-chief (2007–2009) of the *IEEE Transactions on Wireless Communications*. He is a past president of the IEEE Information Theory Society and the IEEE Communications Society.